An Approach for Identifying Author Profiles of Blogs

Chunxia Zhang[1(✉)], Yu Guo[1], Jiayu Wu[1], Shuliang Wang[1], Zhendong Niu[2], and Wen Cheng[3]

[1] School of Software, Beijing Institute of Technology, Beijing, China
{cxzhang,2220160601,2220160656,slwang2011}@bit.edu.cn
[2] School of Computer Science, Beijing Institute of Technology, Beijing, China
zniu@bit.edu.cn
[3] School of Aerospace Engineering, Beijing Institute of Technology, Beijing, China
572705622@qq.com

Abstract. Author profile identification has been an important research problem in the areas of web mining, network public opinion monitoring and social network analysis. The aim of this problem is to identify characteristics or traits of authors of textual information such as blogs, microblogs or reviews in social network platforms or commercial platforms. The technology of author profile identification can be employed into many applications including cyberspace forensics, electronic commerce and information security. In this paper, we propose a hybrid framework or technique to solve the author profile identification problem. In this framework, we design a distributed integrated representation approach of blogs based on Doc2vec and term frequency-inverse document frequency, and apply the convolutional neural network to predict age, gender and education status of authors of blogs. The benefit of our technique is that it predicts three different traits of authors in a uniform way, is an unsupervised method which can learn representation vectors of blog posts based on unlabeled data, and does not need any syntactic and semantic parsing of sentences. Experimental results on blogs show that our approach achieves a promising performance.

Keywords: Author profile identification · Doc2vec · Convolutional neural network · Age prediction · Gender prediction · Education status prediction

1 Introduction

The task of author profile identification or author profiling is to identify characteristics or traits of authors of textual information such as blogs, microblogs, or reviews in social network platforms or commercial platforms. The traits of authors consist of age, gender, location, education status and native language and so on. Author profile identification is an important research problem in the areas of web mining, network public opinion monitoring, social network analysis and opinion mining.

G. Cong et al. (Eds.): ADMA 2017, LNAI 10604, pp. 475–487, 2017.
https://doi.org/10.1007/978-3-319-69179-4_33

The author profile identification technology can be used into many application areas including cyberspace forensics, electronic commerce and information security [1–5]. For example, the author profiling technique can be great helpful to distinguish identity of Internet crimers who may commit network theft and fraud, terrorism, or child predation through social media [1,6]. In addition, the author profiling can be highly beneficial for targeted marketing and advertising, product and service development, product and service review mining [7]. However, it is difficult to fulfill the author profile identification task through manual recognition and detection [6]. Therefore, this paper aims to identify age, gender and educational status of authors of blogs.

As a text genre in social media, blogs have two main characteristics: (1) sentences within blogs maybe contain a lot of non-standard, informal or spoken words, phrases and language usages. For instance, there are abbreviations, internet slangs, or emoticons in texts of blogs. (2) Unlike novels, books or other traditional documents, there are various topics within blog posts, and blog entries are relatively short with personal or subjective thoughts and views.

The key difficulties of solving the author profile identification problem are given as follows: (1) which features are extracted to identify different traits of authors of blogs such as age, gender and education status, and those features are independent of specific topics of blog posts; (2) how to design a uniform generation method of blog representation and a uniform identification method to predict different traits of authors of Internet documents.

In this paper, we propose a hybrid framework or technique to fulfill the author profile identification task. In this framework, we design a distributed integrated representation approach of blogs based on Doc2vec and term frequency-inverse document frequency (TF-IDF), and apply the convolutional neural network to recognize age, gender and education status of authors of blogs. First, we utilize a document representation method based on Doc2vec to generate the distributed representation of blog posts. Second, we construct the representation of blog posts based on TF-IDF. Third, we build the distributed integrated representation of blog posts in terms of the former two kinds of generated representation. Finally, we use the convolutional neural network (CNN) to predict traits of authors of blogs. Experimental results on blogs show that our technique obtains a high performance and outperforms the baseline method.

The main contributions of this paper are given as follows. (1) We propose a distributed integrated representation method for blogs. That method does not rely on any syntactic and semantic parsing of blog posts, and can capture semantic associations between words in sentences and topics of blog posts. Moreover, that representation approach is an unsupervised one which can learn blog post vectors based on unlabeled data. (2) This paper offers a promising technique to identify the age, gender and education status of authors of blogs in a uniform way, that is, it provides a unified pipeline to accomplish the author profile identification task. Experimental results demonstrate that our hybrid technique including Doc2vec, TF-IDF and convolutional neural network outperforms the baseline method, and achieves the higher performance than those of the blog representation method based on TF-IDF or Doc2vec by using decision tree, random forest, and sequential minimal optimization.

The rest of the paper is organized as follows. Section 2 discusses related works about author profile identification. Section 3 presents our approach to solving the author profile identification task. Experimental results are given in Sect. 4. Section 5 concludes the paper and discusses future works.

2 Related Works

The problem of author profile identification or author trait prediction in most works is treated as a two-class or multi-class text classification problem. There have been lots of works on gender identification of authors of blogs, microblogs, news texts, e-mails or PhD theses [1,2,8–16].

Argamon et al. [2] first built style-related features about function words and part-of-speech, and content-based features about the most frequent words to express documents. Further, they used the Bayesian Multinomial Regression to identify age, gender, personality and native language of writers of blogs. Mukherjee and Liu [12] proposed a kind of features about part-of-speech sequence patterns to represent documents, and used support vector machine classification, support vector machine regression and Naive Bayes to identify the gender of blog authors. In addition, the work of Mikros [11] was to predict the gender of authors of Greek blogs. They built features about the most frequent words, character n-grams and stylometric variables, and adopted support vector machine to classify the gender of writers. Furthermore, Ansari et al. [9] constructed three mutual independent features about frequency counter, TF-IDF of tokens and part-of-speech, and then used ZeroR and Naive Bayes to classify gender of blog authors.

Ramnial et al. [8] extracted features about combined-words, words endings, function words, part-of-speech tags and statistics of characters, words, sentences and punctuations as stylometric features of PhD theses, and applied two classifiers of k-nearest neighbour and support vector machine to predict the gender of writers of those PhD theses. Wang et al. [10] first developed two classifiers based on features about user names and microblog texts, respectively. Further, they employed the Bayes rule to integrate two classifiers to recognize the gender of microblog authors. Moreover, Cheng et al. [1] proposed 545 character-based features, word-based features including psycho-linguistic words, syntactic features, structural features and function words including gender-preferential words to represent documents. Three machine learning approaches of support vector machine, Bayesian logistic regression and AdaBoost decision tree were utilized to distinguish the gender of authors of news texts and e-mails.

Relatively, there is a small amount of works on age and education status predication of writers of Internet texts [2,16,17]. Nguyen et al. [17] used a logical and linear regression algorithm to classify Twitter users into three age categories (20−, 20–40, 40+). In addition, Alvarez-Carmona et al. [16] extracted features based on second order attributes and latent semantic analysis to express texts in Twitter, and applied the support vector machine to predict the gender, age and personality of authors.

3 An Approach of Identifying Author Profiles of Blogs

3.1 Problem Formulation

The definition of author profile identification task in this paper is given as follows.

Definition 1 (Author Profile Identification Task). Given a set of authors with known age, gender and education status and their blogs, the author profile identification or author profiling task is to identify the age, gender and education status of authors of anonymous blog posts. In other words, an anonymous blog post is assigned to one of four classes in the set C_{age}, one of two classes in the set C_{gender}, and one of three class in the set $C_{education}$, as shown in (1), respectively.

$$\begin{array}{l} C_{age} = \{25-,\ 26-40,\ 41-60,\ 60+\} \\ C_{gender} = \{male,\ female\} \\ C_{education} = \{postgraduate,\ undergraduate,\ others\} \end{array} \tag{1}$$

3.2 Overview of Our Framework

In general, our author profile identification framework or technique of age, gender and education status consists of four steps, as shown in Fig. 1. (1) Crawling and

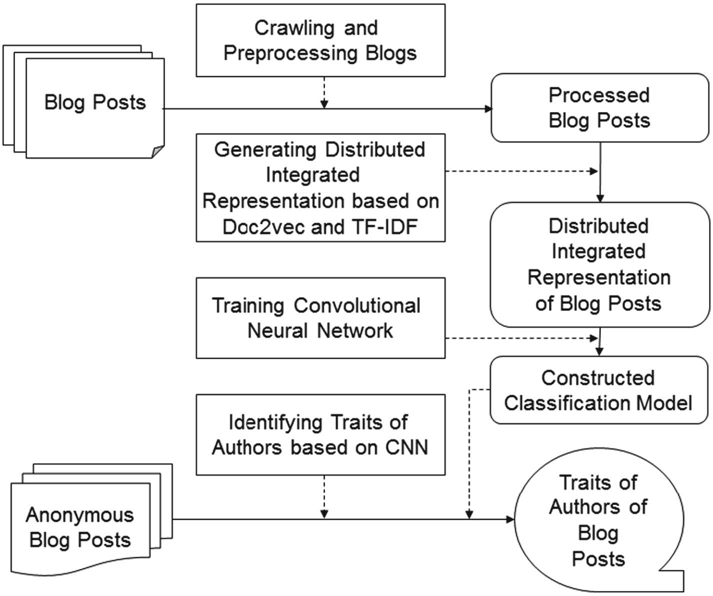

Crawling and Preprocessing Blogs
Blog Posts
Processed Blog Posts
Generating Distributed Integrated Representation based on Doc2vec and TF-IDF
Distributed Integrated Representation of Blog Posts
Training Convolutional Neural Network
Constructed Classification Model
Identifying Traits of Authors based on CNN
Anonymous Blog Posts
Traits of Authors of Blog Posts

Fig. 1. The framework of our author profile identification of blogs

processing blogs to obtain a set of blogs of authors with known age, gender and education status. (2) Building text representation based on Doc2vec and TF-IDF to generate the distributed integrated representation of blog posts. (3) Training the convolutional neural network (CNN) to construct the classification model. (4) Identifying traits of authors of anonymous blog posts based on the trained CNN classification model.

3.3 Generating Distributed Integrated Representation of Blog Posts Based on Doc2vec and TF-IDF

The process of generating vectors of blog posts in this paper is illustrated on the upper part of Fig. 2 [18]. For a blog post, we first learn its distributed representation vector u based on Doc2vec [18]. Second, we build the vector w of each blog post based on TF-IDF values of feature words. Third, we fuse those two vectors to generate the distributed integrated representation of blog posts.

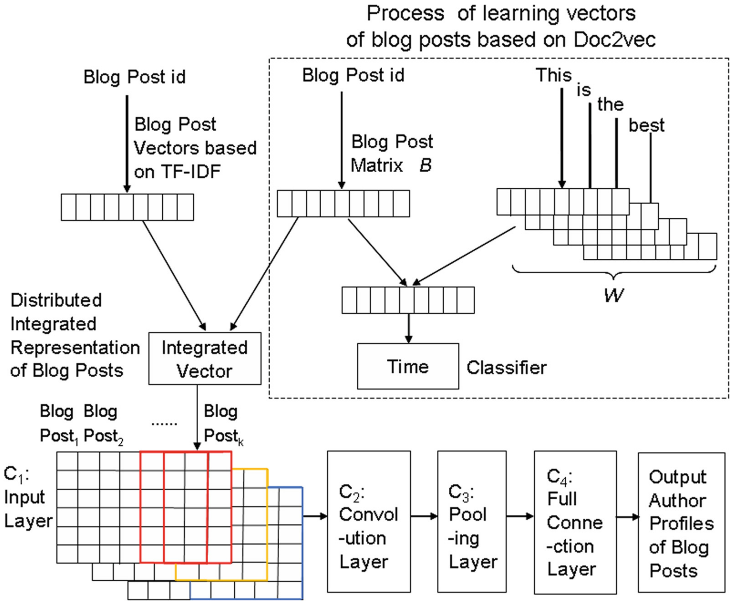

Process of learning vectors of blog posts based on Doc2vec
Blog Post id
Blog Post Vectors based on TF-IDF
Blog Post id
Blog Post Matrix B
This
is
the
best
W
Time
Classifier
Distributed Integrated Representation of Blog Posts
Integrated Vector
Blog Post$_1$
Blog Post$_2$
......
Blog Post$_k$
C$_1$: Input Layer
C$_2$: Convol -ution Layer
C$_3$: Pool -ing Layer
C$_4$: Full Conne -ction Layer
Output Author Profiles of Blog Posts

Fig. 2. The process of our approach to identify author profiles of blogs.

In Fig. 2, each blog post within blogs of an author is viewed as a document. Every blog post and every word in blog posts corresponds to a unique vector. The matrix B in Fig. 2 is composed of vectors of different blog posts, where a column in B denotes a blog post. In addition, the matrix W in Fig. 2 is made up of vectors of different words, where a column in W expresses a word.

The core idea of Doc2vec is to extend word embedding to document embedding, that is, it is an extension of Word2vec [18–21]. Actually, document embedding can generate vectors of sentences, paragraphs, or documents. Word2vec aims to build the word embedding based on contexts of words in sentences. In other words, the goal of word2vec is to construct a vector for each word according to its contextual words within sentences. The difference between Doc2vec and Word2vec is that a document vector is introduced into word2vec to capture the topic of the document or the missing information of the current context [18].

The reasons that we employed Doc2vec and TF-IDF to generate distributed vectors of blog posts are given as follows [18–21]. (1) Doc2vec is an unsupervised approach which can generate vector representations of different length of blog posts. (2) The distributed vector representations of blog posts are intended to capture semantic associations between words in sentences. (3) The blog representation based on TF-IDF highlights the differences between stylistic word features of blogs with different traits of authors. (4) We do not need any syntactic and semantic parsing on sentences in blogs, since Doc2vec learns document vectors based on unlabeled data.

3.4 An Identification Algorithm of Author Profiles

When a blog post has been expressed as a distributed integrated vector, we employ the convolutional neural network (CNN) to predict profiles or traits of authors of blogs, as shown in Fig. 2. The reason that we use the CNN model is that it has the good fault tolerance ability, self learning ability and local sensing ability, and decreases the complexity of the neural network models [22,23].

In the input layer C_1 in Fig. 2, we build the distributed integrated vector representation of each blog post based on Doc2vec and TF-IDF [22,23]. Through the input layer, the blog posts of an author is mapped into a $m \times n$ matrix, where m is the number of dimensions of the vector of each blog post, n is the number of blog posts of this author.

The layer C_2 in Fig. 2 is a convolution layer. The higher abstract features of blog posts are extracted through the convolution layer [22,23]. For instance, we can generate a feature d_i based on a window of blog posts, as shown in (2).

$$d_i = f(W * v_{i:i+h-1} + b) \tag{2}$$

Here, $v_{i:i+h-1}$ is the concatenation of $v_i, v_{i+1}, ..., v_{i+h-1}$, v_i is the vector of the i-th blog post, the filter W and b are parameters of the convolutional kernel, and f is a non-linear function. Further, we build a feature map shown in (3) by executing convolution on all possible windows of blog posts.

$$d = (d_1,\, d_2,\, ...,\, d_{n-h+1}) \tag{3}$$

In the pooling layer C_3 in Fig. 2, we obtain distinct features v_{max} of blog posts, shown in (4), by performing the max-pooling operation over the feature map [22,23].

$$v_{max} = max\{v_1,\, v_2,\, ...,\, v_n\} \tag{4}$$

Further, those distinct features are input into the full connection layer C_4 in Fig. 2 with dropout and softmax. And we can get the probability distribution over different trait classes for blog posts [22,23]. Now, we give our author profile identification algorithm in Algorithm 1.

Algorithm 1. *Identifying Author Profiles of Blogs*

```
Input: The blog posts $B$ of anonymous authors.
Output: The age, gender and education status of authors of blog posts in $B$.
for $b_i \in B$, $i$=1, 2,..., $n$ do
    Use a word segmentation tool to separate words in sentences in the blog post $b_i$.
    Generate the distributed vector representation $u_i$ of $b_i$ based on Doc2vec.
    Build the vector $w_i$ of $b_i$ based on TF-IDF.
    Construct the distributed integrated representation $(u_i,w_i)$ of $b_i$.
end for
Classify the author of each blog post into one class in $C_{age}$, $C_{gender}$ and $C_{education}$
    based on the convolutional neural network(CNN) in a uniform pipeline.
```

4 Experiments

4.1 Experimental Results

In our experiments, we downloaded Chinese blogs of fifty-two famous persons with about 8700 blog posts from the website "http://www.sina.com.cn/" for evaluation of our technique. That website is one of the biggest portal sites in China, and Sina blog is one of the most popular blog channels in China. The number of blog posts of each author is ranging from 15 to 675. In order to build datasets as balanced as possible, we selected three different sets of downloaded blog posts as the datasets for gender, education status and age identification. The dataset used in the experiment of gender identification of authors of blog posts include more than 5100 blog posts of twenty-four persons, the dataset for education status identification consists of more than 4900 blog posts of thirty-four persons, and the dataset for age identification contains more than 5200 blog posts of thirty-two persons. The ten-fold cross validation is used to evaluate the performance of our author profiles identification technique. The baseline method is one which builds vectors of blog posts based on TF-IDF and employs the decision tree J48 to deal with the author profile identification task.

Table 1 gives the identification accuracy of gender, education status and age of authors of blogs by using our proposed technique in this paper and nine approaches which are combinations of three kinds of representation methods of blog posts and three types of prediction methods of authors' traits. Those three kinds of representation methods of blog posts include representations based on TF-IDF, Doc2vec, and the integration of TF-IDF and Doc2vec. Those three types of prediction methods of authors' traits consist of the decision tree J48 (DT), random forest (RF), and sequential minimal optimization (SMO). For example, the fifth method "Doc2vec ⊕ SMO" in Table 1 means one which constructs vectors of blog posts according to Doc2vec and utilizes SMO to recognize author profiles of authors of blog posts. Here, SMO is an optimization method for training support vector machines [24]. In Table 1, the dimensions of representation vectors of blog posts based on TF-IDF for gender, education status and age identification is 501, 503, 503, respectively. And the dimensions of representation vectors of blog posts based on Doc2vec in Table 1 is 1000.

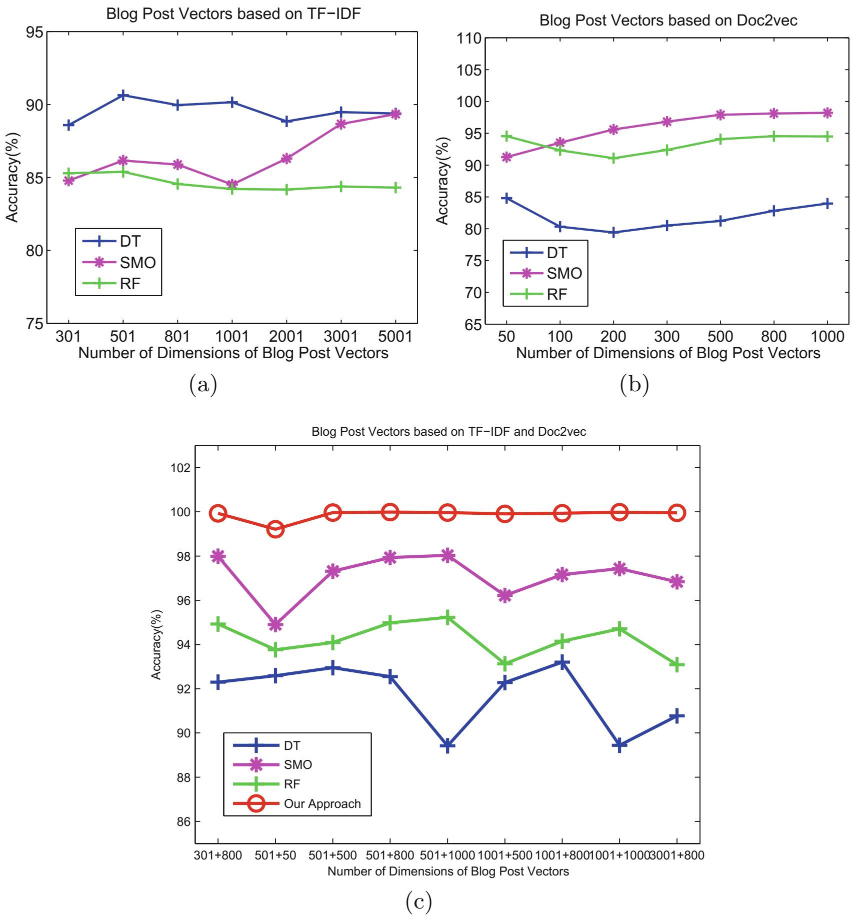
Table 1. The identification accuracy of gender, education status and age of authors of blogs

NO.	Accuracy (%)	Gender	Education status	Age
1	Baseline (TF-IDF ⊕ DT)	90.6371	87.0040	90.0661
2	TF-IDF ⊕ SMO	86.1583	82.4899	89.0652
3	TF-IDF ⊕ RF	85.3861	83.5830	87.6865
4	Doc2vec ⊕ DT	83.9575	61.8421	66.0812
5	Doc2vec ⊕ SMO	98.2046	90.0607	96.3362
6	Doc2vec ⊕ RF	94.4981	83.1781	94.0321
7	Doc2vec ⊕ TF-IDF ⊕ DT	89.4208	82.2672	91.1237
8	Doc2vec ⊕ TF-IDF ⊕ SMO	98.0309	97.6721	98.3947
9	Doc2vec ⊕ TF-IDF ⊕ RF	95.2317	94.1498	95.5052
10	Our Approach (Doc2vec ⊕ TF-IDF ⊕ CNN)	99.9676	97.7473	96.3197

The experimental results in Table 1 indicate that the following facts. (1) our hybrid technique achieves the highest identification accuracy of gender and education status than those of other nine kinds of approaches, and obtains the higher identification accuracy of age than those of other seven types of methods (i.e., the 1st, 2nd, 3rd, 4th, 6th, 7th and 9th methods). (2) Our distributed integrated representation of TF-IDF and Doc2vec gets the higher accuracy than those of blog post representation based on TF-IDF or Doc2vec by using SMO and RF for age and education status identification.

4.2 Parameters Analysis

We analyse the performance influence of different dimensions of representation vectors of blog posts. The vector of blog posts based on Doc2vec are set as 50, 100, 200, 300, 500, 800 and 1000 dimensions. The vector of blog posts based on TF-IDF are set up from 301 to 5003 dimensions. Figures 3, 4 and 5 give the accuracy curves of our hybrid technique and the accuracy curves of DT, RF and



(a)

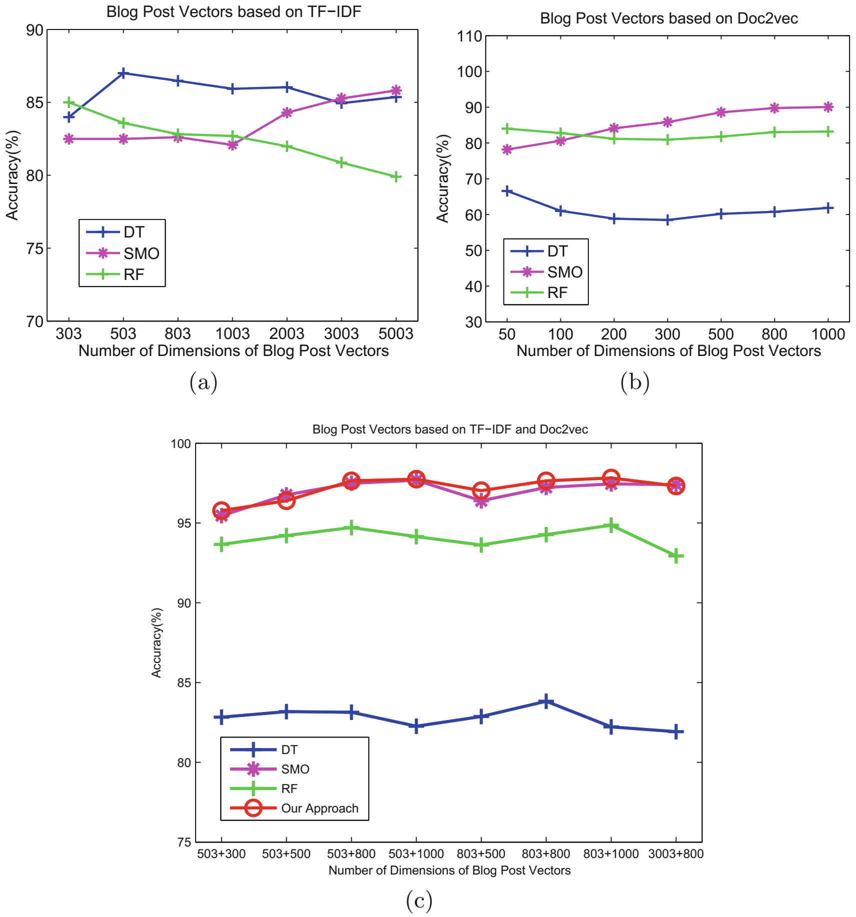
(b)

(c)

Fig. 3. The gender identification accuracy of different methods. (a) The accuracies with different dimensions of the blog post vectors obtained by TF-IDF; (b) The accuracies with different dimensions of the blog post vectors obtained by Doc2vec; (c) The accuracies achieved by combing together the features obtained respectively by TF-IDF and Doc2vec with different dimensions of the blog post vectors. For example, "301 + 800" means that the features of 301 dimensions obtained by TF-IDF and the features of 800 dimensions obtained by Doc2vec are combined together.

SMO by representing blog posts based on TF-IDF, Dov2vec and the distributed integrated representation of TF-IDF and Doc2vec, respectively.

We can see the following facts from Figs. 3, 4 and 5. (1) The gender prediction accuracy of SMO is higher than those of DT and RF based on two kinds of blog representation methods in most cases of Fig. 3(b) and (c). Those two kinds of blog representation methods include representations based on Doc2vec and the integration of TF-IDF and Doc2vec. This fact holds for the education status identification in most cases of Fig. 4(b) and (c) and for the age identification



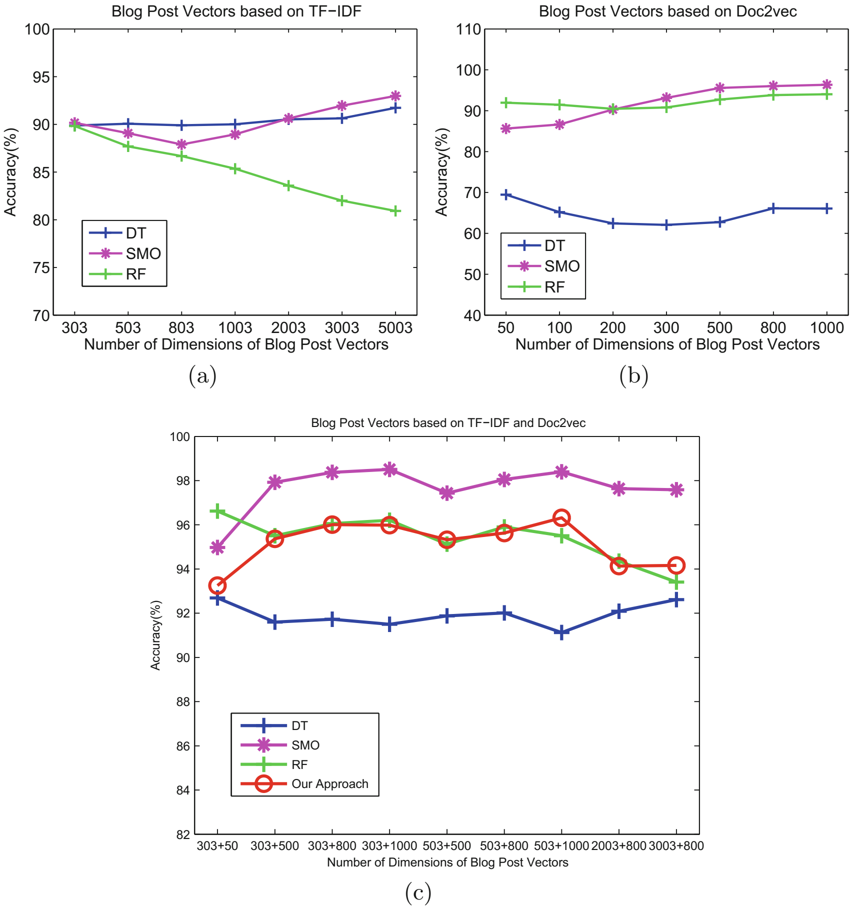
(a)

(b)

(c)

Fig. 4. The education status identification accuracy of different methods. (a) The accuracies with different dimensions of the blog post vectors obtained by TF-IDF; (b) The accuracies with different dimensions of the blog post vectors obtained by Doc2vec; (c) The accuracies achieved by combing together the features obtained respectively by TF-IDF and Doc2vec with different dimensions of the blog post vectors. For instance, "503 + 300" means that the features of 503 dimensions obtained by TF-IDF and the features of 300 dimensions obtained by Doc2vec are combined together.

in most cases of Fig. 5(b) and (c). (2) In general, our hybrid technique obtains the highest accuracy of the gender and education status identification in Figs. 3 and 4. Specifically, the identification accuracy of gender of blog authors by using our approach gets about 99.987%, while the identification accuracy of education status with our algorithm achieves about 97.826%.



(a)

(b)

(c)

Fig. 5. The age identification accuracy of different methods. (a) The accuracies obtained by TF-IDF with different dimensions of the blog post vectors; (b) The accuracies obtained by Doc2vec approach with different dimensions of the blog post vectors; (c) The accuracies achieved by combing together the features obtained respectively by TF-IDF and Doc2vec approach with different dimensions of the blog post vectors. For example, "$303 + 50$" means that the features of 303 dimensions obtained by TF-IDF and the features of 50 dimensions obtained by Doc2vec are combined together.

5 Conclusion

More and more efforts have been paid on author profile identification or author profiling from Internet documents such as microblogs, blogs and product reviews. The author profile identification technology has wide applications containing cyberspace forensics, electronic commerce, information recommendation and information security. In this paper, a hybrid framework or technique is proposed to solve the author profile identification task. Specifically, in this framework, we build a distributed integrated representation of blog posts based on Doc2vec and term frequency-inverse document frequency, and employ the convolutional neural network to predict age, gender and education status of authors of blogs. The traits of our technique is that it is an unsupervised approach which can learn distributed vectors of blog posts based on unlabeled data, does not require any syntactic and semantic parsing of blog posts, and predict three different traits of authors in a uniform pipeline. Experimental results on blogs indicate that our approach is valid. In the future, we will design methods to identify location and personality of authors of textual information in social media.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (NO. 61672098) and Frontier and Interdisciplinary Innovation Program of Beijing Institute of Technology (NO. 3080012291701).

References

1. Cheng, N., Chandramouli, R., Subbalakshmi, K.P.: Author gender identification from text. Digital Invest. **8**(1), 78–88 (2011)
2. Argamon, S., Koppel, M., Pennebaker, J., et al.: Automatically profiling the author of an anonymous text. Commun. ACM **52**(2), 119–123 (2009)
3. Rangel, F., Rosso, C., Fabio, M., et al.: Overview of the 3rd author profiling task at PAN 2015. In: CLEF 2015 Evaluation Labs and Workshop Working Notes Papers, pp. 1–8 (2015)
4. Op Vollenbroek, M.B., Carlotto, T., Kreutz, T., et al.: GronUP: Groningen user profiling notebook for PAN at CLEF 2016. In: CLEF 2016 Evaluation Labs and Workshop Working Notes Papers (2016)
5. Wang, L.: Author profiling. Master's Thesis. Beijing Institute of Technology, Beijing, China (2013)
6. Peersman, C., Daelemans, W., Van Vaerenbergh, L.: Predicting age and gender in online social networks. In: 3rd International Workshop on Search and Mining User-Generated Contents, pp. 37–44 (2011)
7. Zhang, C., Zhang, P.: Predicting gender from blog posts (2010). http://web.stanford.edu/~pyzhang/papers/gender_prediction.pdf
8. Ramnial, H., Panchoo, S., Pudaruth, S.: Gender profiling from PhD theses using k-nearest neighbour and sequential minimal optimisation. In: Berretti, S., Thampi, S.M., Dasgupta, S. (eds.) Intelligent Systems Technologies and Applications. AISC, vol. 385, pp. 369–377. Springer, Cham (2016). doi:10.1007/978-3-319-23258-4_32
9. Ansari, Y.Z., Azad, S.A., AKhtar, H., et al.: Gender classification of blog authors. Int. J. Sustain. Dev. Green Econ. (2013). Special Issue
10. Wang, J., Li, S., Huang, L.: User gender classification in Chinese microblog. J. Chin. Inf. Process. **28**(6), 150–155 (2014)
11. Mikros, G.K.: Authorship attribution and gender identification in Greek blogs. Meth. Appl. Quant. Linguist. **21**, 21–32 (2012)
12. Mukherjee, A., Liu, B.: Improving gender classification of blog authors. In: The Conference on Empirical Methods in Natural Language Processing, pp. 207–217 (2010)
13. Miller, Z., Dickinson, B., Hu, W.: Gender prediction on twitter using stream algorithms with N-gram character features. Int. J. Intell. Sci. **2**(4), 143–148 (2012)

14. Wang, F.: A study on gender classification of blog authors. Master's Thesis. Beijing Jiaotong University. Beijing, China (2012)
15. Yang, J.: Research on gender recognition technology of Chinese e-mail authors based on SVM. Master's Thesis. Hebei Agricultural University, Hebei, China (2007)
16. Alvarez-Carmona, M., Lopez-Monroy, P., et al.: INAOE's participation at PAN'15: author profiling task-notebook for PAN at CLEF 2015. In: CLEF 2015 Evaluation Labs and Workshop Working Notes Papers (2015)
17. Nguyen, D., Gravel, R., Trieschnigg, D., et al.: How old do you think I am? A study of language and age in Twitter. In: 7th International AAAI Conference on Weblogs and Social Media, pp. 1–10 (2013)
18. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: 31st International Conference on Machine Learning, pp. 1188–1196 (2014)
19. Lau, J.H., Baldwin, T.: An empirical evaluation of Doc2vec with practical insights into document embedding generation (2016). https://arxiv.org/pdf/1607.05368.pdf
20. Word embedding. https://en.wikipedia.org/wiki/Word_embedding
21. Word2vec. https://en.wikipedia.org/wiki/Word2vec
22. Kim, Y.: Convolutional Neural Networks for Sentence Classification (2014). http://www.aclweb.org/anthology/D14-1181
23. Hu, B., Lu, Z., Li, H., et al.: Convolutional neural network architectures for matching natural language sentences(2014). http://www.hangli-hl.com/uploads/3/1/6/8/3168008/hu-etal-nips2014.pdf
24. Sequential minimal optimization. https://en.wikipedia.org/wiki/Sequential_minimal_optimization